LAGRANGE MULTIPLIERS

14

We know that a "nice" function on a compact manifold (a closed curve, soap bubble, etc.) has a maximum and minimum value. But so far our only tool for finding them rests on parametrizing the manifold, like $t \mapsto \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}$ for the unit circle. This may not always be possible, though more often, it is just inconvenient and makes the problem harder to solve. Fortunately, there is a better approach, and that is the subject of this lecture.

§1. One Lagrange multiplier

A little less heuristically, if $\vec{r}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$ parametrizes $g\begin{pmatrix} x \\ y \end{pmatrix} = 0$, then the function f is maximized on $g=0$ where

$$0 = \frac{d}{dt} f(\vec{r}(t)) = \vec{\nabla} f(\vec{r}(t)) \cdot \vec{r}'(t)$$

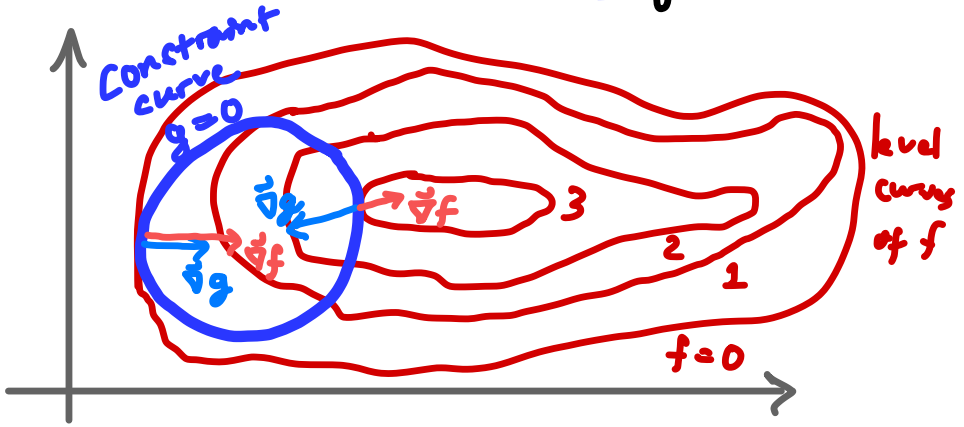
for $t =$ some t_0; and then $\vec{\nabla} f(\vec{r}(t_0)) \perp \vec{r}'(t_0)$.

But $\vec{\nabla} g(\vec{r}(t_0)) \perp \vec{r}'(t_0)$ as well (since gradient is normal to $g=0$, and $\vec{r}'(t_0)$ tangent to it), so $\vec{\nabla} f(\vec{r}(t_0)) \parallel \vec{\nabla} g(\vec{r}(t_0))$.

(they are parallel)

§1. One Lagrange multiplier

Consider first the case where we want to optimize $f\begin{pmatrix} x \\ y \end{pmatrix}$ subject to $g\begin{pmatrix} x \\ y \end{pmatrix} = 0$:



Taking it as geometrically evident that the level curve $f=k$ with greatest possible k intersecting the constraint curve is tangent to it at the intersection point, we conclude that their normal vectors are parallel, and

$$\vec{\nabla} f = \lambda \vec{\nabla} g$$

at $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ on $g=0$ where f is $\begin{cases} \text{maximized} \\ \text{minimized.} \end{cases}$

A little less heuristically, if $\vec{r}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$ parametrizes $g\begin{pmatrix} x \\ y \end{pmatrix} = 0$, then the function f is maximized on $g=0$ where

$$0 = \frac{d}{dt} f(\vec{r}(t)) = \vec{\nabla} f(\vec{r}(t)) \cdot \vec{r}'(t)$$

for $t =$ some t_0; and then $\vec{\nabla} f(\vec{r}(t_0)) \perp \vec{r}'(t_0)$.

But $\vec{\nabla} g(\vec{r}(t_0)) \perp \vec{r}'(t_0)$ as well (since gradient is normal to $g=0$, and $\vec{r}'(t_0)$ tangent to it), so $\vec{\nabla} f(\vec{r}(t_0)) \parallel \vec{\nabla} g(\vec{r}(t_0))$.

(they are parallel)

This approach to solving constrained extremum problems generalizes to n variables, yielding the algorithm:

1) put constraint in form $g(\vec{x}) = 0$
2) set $\vec{\nabla} f = \lambda \vec{\nabla} g$
3) solve the resulting $n+1$ equations for $\vec{x}, \lambda$
4) evaluate f at each solution $\vec{x}$

Exercise: Find the maximum volume of a box inscribed in the ellipsoid $\frac{x^2}{4} + \frac{y^2}{2} + z^2 = 1$.

(1) $f = 8xyz$, $g = x^2 + 2y^2 + 4z^2 - 4$

(2) $8\begin{pmatrix} yz \\ xz \\ xy \end{pmatrix} = \lambda \begin{pmatrix} 2x \\ 4y \\ 8z \end{pmatrix}$

(3)
A: $\lambda x = 4yz \overset{B}{=} 4\left(\frac{2}{\lambda} xz\right) z = \frac{8}{\lambda} xz^2$
B: $\lambda y = 2xz \overset{C}{=} 2x\left(\frac{1}{\lambda} xy\right) = \frac{2}{\lambda} x^2 y$
C: $\lambda z = xy \overset{A}{=} \left(\frac{4}{\lambda} yz\right) y = \frac{4}{\lambda} y^2 z$
D: $x^2 + 2y^2 + 4z^2 = 4$

$z^2 = \frac{\lambda^2}{8}$ $\quad x^2 = \frac{\lambda^2}{2}$ $\quad y^2 = \frac{\lambda^2}{4}$

$z = \pm \frac{\lambda}{2\sqrt{2}}$ $\quad x = \pm \frac{\lambda}{\sqrt{2}}$ $\quad y = \pm \frac{\lambda}{2}$

↳ plug in to D: $\frac{\lambda^2}{2} + \frac{\lambda^2}{2} + \frac{\lambda^2}{2} = 4$

$\Rightarrow \lambda^2 = \frac{8}{3} \Rightarrow \lambda = \pm \frac{2\sqrt{2}}{\sqrt{3}}$

$\Rightarrow (x, y, z) = \left(\pm \frac{2}{\sqrt{3}}, \pm \frac{\sqrt{2}}{\sqrt{3}}, \pm \frac{1}{\sqrt{3}}\right)$ choose "+"

(4) evaluate $f(\cdots) = 8 \cdot \frac{2}{\sqrt{3}} \cdot \frac{\sqrt{2}}{\sqrt{3}} \frac{1}{\sqrt{3}} = \frac{16\sqrt{2}}{3\sqrt{3}}$

max. volume

§2. Several Lagrange multipliers

To find all candidates for extrema of $f\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ subject to m constraints $\begin{cases} g_1(\vec{x}) = 0 \\ \vdots \\ g_m(\vec{x}) = 0 \end{cases}$

we need to solve the (nonlinear) system

$$\begin{cases} g_1(\vec{x}) = 0 \\ \vdots \\ g_m(\vec{x}) = 0 \\ \vec{\nabla} f(\vec{x}) = \lambda_1 \vec{\nabla} g_1(\vec{x}) + \cdots + \lambda_m \vec{\nabla} g_m(\vec{x}) \end{cases}$$

$n+m$ equations

↳ this is really n equations:

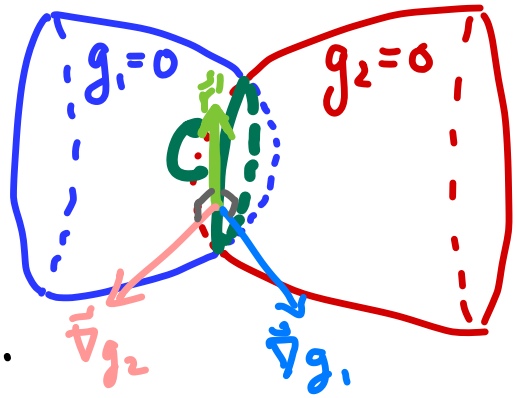
$$\left.\begin{aligned} \partial f/\partial x_1 &= \lambda_1 \partial g_1/\partial x_1 + \cdots + \lambda_m \partial g_m/\partial x_1 \\ &\vdots \\ \partial f/\partial x_n &= \lambda_1 \partial g_1/\partial x_n + \cdots + \lambda_m \partial g_m/\partial x_n \end{aligned}\right\}$$

So altogether we have $n+m$ equations in $n+m$ variables $\lambda_1, \ldots, \lambda_m, x_1, \ldots, x_n$.

One expects the solution set "in general" is a finite set of points, which will include our desired maximum & minimum.

But **why should the method work?**

IDEA Say $n=3$, $m=2$ and we are trying to max/minimize $f\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ on $C = \{g_1 = 0\} \cap \{g_2 = 0\}$.



If $\vec{r}(t)$ parametrizes C, then $\vec{r}'(t)$ is tangent to C, hence tangent to $g_1=0$ & $g_2=0$.

Thus $\vec{r}' \perp \vec{\nabla} g_1$ & $\vec{\nabla} g_2$ everywhere on C.

At an extremum of f on C (at $\vec{r}_0 = \vec{r}(t_0)$),

$$0 = \frac{d}{dt} f(\vec{r}(t))\Big|_{t=t_0} = \vec{\nabla} f(\vec{r}_0) \cdot \vec{r}'(t_0)$$

$$\Rightarrow \vec{r}' \perp \vec{\nabla} f.$$

§2. Several Lagrange multipliers

To find all candidates for extrema of $f\begin{pmatrix}x_1\\ \vdots \\ x_n\end{pmatrix}$ subject to m constraints $\begin{cases} g_1(\vec{x}) = 0 \\ \vdots \\ g_m(\vec{x}) = 0 \end{cases}$

we need to solve the (nonlinear) system

$$(*)\begin{cases} g_1(\vec{x}) = 0 \\ \vdots \\ g_m(\vec{x}) = 0 \\ \vec{\nabla} f(\vec{x}) = \lambda_1 \vec{\nabla} g_1(\vec{x}) + \cdots + \lambda_m \vec{\nabla} g_m(\vec{x}) \end{cases}$$

$n+m$ equations

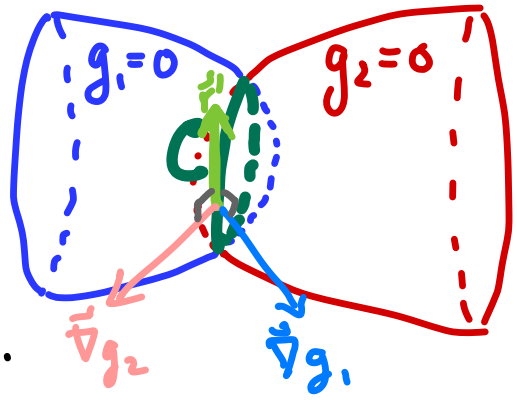
↳ this is really n equations:

$$\left.\begin{aligned} \partial f/\partial x_1 &= \lambda_1 \partial g_1/\partial x_1 + \cdots + \lambda_m \partial g_m/\partial x_1 \\ &\vdots \\ \partial f/\partial x_n &= \lambda_1 \partial g_1/\partial x_n + \cdots + \lambda_m \partial g_m/\partial x_n \end{aligned}\right\}$$

So to complete the method, we should throw into the critical set, along with solutions to $(*)$, all singular points of $g_1 = \cdots = g_m = 0$ — points where $D\vec{g}$ is not of maximal rank.

But why should the method work?

IDEA Say $n=3$, $m=2$ and we are trying to max/minimize $f\begin{pmatrix}x\\ y\\ z\end{pmatrix}$ on $C = \{g_1 = 0\} \cap \{g_2 = 0\}$.



If $\vec{r}(t)$ parametrizes C, then $\vec{r}'(t)$ is tangent to C, hence tangent to $g_1=0$ & $g_2=0$.

Thus $\vec{r}' \perp \vec{\nabla} g_1$ & $\vec{\nabla} g_2$ everywhere on C.

At an extremum of f on C (at $\vec{r}_0 = \vec{r}(t_0)$),

$$0 = \frac{d}{dt} f(\vec{r}(t))\Big|_{t=t_0} = \vec{\nabla} f(\vec{r}_0) \cdot \vec{r}'(t_0)$$

$\Rightarrow \vec{r}' \perp \vec{\nabla} f$.

! this will happen as long as they are independent, i.e. C is smooth at $\vec{r}_0$.

Therefore, if $\vec{\nabla} g_1(\vec{r}_0)$ & $\vec{\nabla} g_2(\vec{r}_0)$ span the subspace $\perp$ to $\vec{r}'(t_0)$, we have

$$\vec{\nabla} f(\vec{r}_0) = \lambda_1 \vec{\nabla} g_1(\vec{r}_0) + \lambda_2 \vec{\nabla} g_2(\vec{r}_0)$$

for some $\lambda_1, \lambda_2 \in \mathbb{R}$.

§2. Several Lagrange multipliers

To find all candidates for extrema of $f\begin{pmatrix}x_1\\ \vdots \\ x_n\end{pmatrix}$ subject to m constraints $\begin{cases} g_1(\vec{x}) = 0 \\ \vdots \\ g_m(\vec{x}) = 0 \end{cases}$

we need to solve the (nonlinear) system

$$(*)\begin{cases} g_1(\vec{x}) = 0 \\ \vdots \\ g_m(\vec{x}) = 0 \\ \vec{\nabla} f(\vec{x}) = \lambda_1 \vec{\nabla} g_1(\vec{x}) + \cdots + \lambda_m \vec{\nabla} g_m(\vec{x}) \end{cases}$$

$n+m$ equations

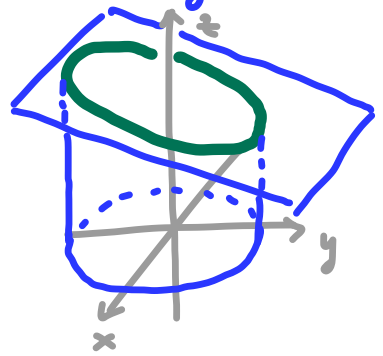
↳ this is really n equations:

$$\left.\begin{aligned} \partial f/\partial x_1 &= \lambda_1 \partial g_1/\partial x_1 + \cdots + \lambda_m \partial g_m/\partial x_1 \\ &\vdots \qquad\qquad \vdots \\ \partial f/\partial x_n &= \lambda_1 \partial g_1/\partial x_n + \cdots + \lambda_m \partial g_m/\partial x_n \end{aligned}\right\}$$

So to complete the method, we should throw into the critical set, along with solutions to $(*)$, all singular points of $g_1 = \cdots = g_m = 0$ — points where Dg is not of maximal rank.

Exercise: Optimize $f\begin{pmatrix}x\\ y\\ z\end{pmatrix} = x + 2y + 3z$

on the ellipse that is the intersection of the cylinder $x^2+y^2=2$ and the plane $y+z=1$.



$g_1 = x^2+y^2-2$, $g_2 = y+z-1$

$$\begin{pmatrix}1\\2\\3\end{pmatrix} = \vec{\nabla} f = \lambda_1 \vec{\nabla} g_1 + \lambda_2 \vec{\nabla} g_2 = \begin{pmatrix} 2\lambda_1 x \\ 2\lambda_1 y + \lambda_2 \\ \lambda_2 \end{pmatrix}$$

$$\rightsquigarrow \begin{cases} x = 1/2\lambda_1 \\ y = 1/\lambda_1 - \lambda_2/2\lambda_1 = -1/2\lambda_1 \\ \lambda_2 = 3 \end{cases}$$

$$\left(\frac{1}{2\lambda_1}\right)^2 + \left(\frac{-1}{2\lambda_1}\right)^2 = 2$$

$$\Rightarrow \lambda_1 = \pm\frac{1}{2}$$

2 cases

- $\lambda_1 = +\frac{1}{2} \Rightarrow \begin{pmatrix}x\\y\\z\end{pmatrix} = \begin{pmatrix}1\\-1\\2\end{pmatrix} \Rightarrow f = 5$ MAX
- $\lambda_1 = -\frac{1}{2} \Rightarrow \begin{pmatrix}x\\y\\z\end{pmatrix} = \begin{pmatrix}-1\\1\\0\end{pmatrix} \Rightarrow f = -1$ MIN

Lagrange multiplier theorem:

$M = \{g_1 = \cdots = g_m = 0\} \subset \mathbb{R}^n$ manifold

$\vec{a} \in M$ (local) extremum of $f : \mathbb{R}^n \to \mathbb{R}$ on M

$$\Rightarrow \vec{\nabla} f(\vec{a}) = \sum_{i=1}^{m} \lambda_i \vec{\nabla} g_i(\vec{a}) \quad \text{for some } \lambda_i \in \mathbb{R}.$$

Proof: By the yet-to-be-proved Implicit Function Theorem, there exists a local parametrization of M

$$\Phi : U \to M$$

sending $\tilde{a} \longmapsto \vec{a}$

where $U \subset \mathbb{R}^{n-m}$ is an open set.

Now: $\vec{a}$ = local extremum of f on M

$\Downarrow$

$\tilde{a}$ = local extremum of $f \circ \Phi$

$\Downarrow$

$\tilde{a}$ = stationary point of $f \circ \Phi$

$\Downarrow$

$$0 = D(f \circ \Phi)(\tilde{a})$$

$\Downarrow$ Chain rule

$$0 = Df(\vec{a}) \circ D\Phi(\tilde{a})$$

$\Downarrow$

$$N(Df(\vec{a})) \supset \underbrace{C(D\Phi(\tilde{a}))}_{\dim = n-m} = \underbrace{N(D\vec{g}(\vec{a}))}_{\dim = n-m} \quad (R+N)$$

Also: $\vec{g} \circ \Phi = 0$

$\Downarrow$

$D(\vec{g} \circ \Phi) = 0$

$\Downarrow$

$D\vec{g}(\vec{a}) \circ D\Phi(\tilde{a}) = 0$

$\Downarrow$ rank = m

$\Downarrow$

$$N(Df(\vec{a})) \supset N(D\vec{g}(\vec{a}))$$

$\perp \Downarrow$ (take orthogonal complements)

$$R(Df(\vec{a})) \subset R(D\vec{g}(\vec{a}))$$

row space $=$ $\qquad$ row space $=$

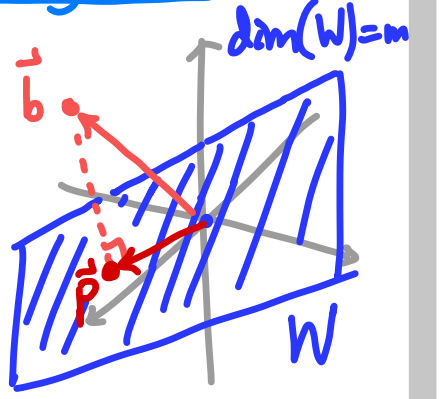
$\operatorname{span}\{\vec{\nabla} f(\vec{a})\}$ $\qquad$ $\operatorname{span}\{\vec{\nabla} g_1(\vec{a}), \ldots, \vec{\nabla} g_m(\vec{a})\}$

$\Downarrow$

$$\vec{\nabla} f(\vec{a}) \in \operatorname{span}\{\vec{\nabla} g_1(\vec{a}), \ldots, \vec{\nabla} g_m(\vec{a})\}.$$

□

§3. Application to projections



Given $\vec{b} \in \mathbb{R}^n$ and a subspace $W \subset \mathbb{R}^n$, how do we find the closest point $\vec{p}$ to $\vec{b}$ on W? We could try to treat it as a constrained extremum problem, and use Lagrange multipliers, but then what are f and the g_i?

For the constraint equations, we should write W as the null-space of a $(n-m) \times n$ matrix B, whose rows $\vec{B}_i$ $(i=1,\dots,n-m)$ yield the $n-m$ equations

$$0 = g_i(\vec{x}) := {}^t\vec{B}_i \cdot \vec{x}$$

cutting W out, with $\vec{\nabla} g_i = {}^t\vec{B}_i$.

For the function to be minimized, we take

$$f(\vec{x}) = \|\vec{x} - \vec{b}\|^2,$$

with $\vec{\nabla} f = 2(\vec{x} - \vec{b})$. So the Lagrange equations read

$$\vec{x} - \vec{b} = \lambda_1 {}^t\vec{B}_1 + \dots + \lambda_{n-m} {}^t\vec{B}_{n-m}$$
$$= {}^tB \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_{n-m} \end{pmatrix} = {}^tB\,\vec{\lambda}.$$

Multiplying both sides by B and using the constraint equation $B\vec{x} = \vec{0}$ gives

$$-B\vec{b} = B\vec{x} - B\vec{b} = B\,{}^tB\,\vec{\lambda}.$$

$W = \mathrm{Nul}(B)$, $\vec{b} \in \mathbb{R}^n$,
want $\vec{p} \in W$ minimizing $\|\vec{p} - \vec{b}\|$.

So multiplying both sides of (**) by $(B^tB)^{-1}$ gives

$$-(B^tB)^{-1}B\vec{b} = \vec{\lambda},$$

and substituting this in (*) yields

$$\boxed{\vec{p} = \vec{b} - {}^tB(B^tB)^{-1}B\vec{b}}$$

which is our projection formula.

For the function to be minimized, we take

$$f(\vec{x}) = \|\vec{x} - \vec{b}\|^2,$$

with $\vec{\nabla} f = 2(\vec{x} - \vec{b})$. So the Lagrange equations read

$$\vec{p} - \vec{b} = \vec{x} - \vec{b} = \lambda_1 {}^t\vec{B}_1 + \cdots + \lambda_{n-m} {}^t\vec{B}_{n-m} = {}^tB \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_{n-m} \end{pmatrix} = {}^tB\vec{\lambda}. \quad (*)$$

Multiplying both sides by B and using the constraint equation $B\vec{x} = \vec{0}$ gives

$$-B\vec{b} = B\vec{x} - B\vec{b} = B\,{}^tB\vec{\lambda}. \quad (**)$$

I claim that B^tB is invertible:

- B has rank $n-m$ (b/c nullity$(B) = \dim W = m$)
  $\Rightarrow {}^tB$ has rank $n-m$ $\Rightarrow$ nullity$({}^tB) = 0$
- if $B\,{}^tB\vec{y} = \vec{0}$, then
  $0 = {}^t\vec{y}\,B\,{}^tB\vec{y} = {}^t({}^tB\vec{y})\,{}^tB\vec{y} = \|{}^tB\vec{y}\|^2$
  $\Rightarrow {}^tB\vec{y} = \vec{0} \Rightarrow \vec{y} = \vec{0}$.

$W = \mathrm{Nul}(B)$, $\vec{b} \in \mathbb{R}^n$,

want $\vec{p} \in W$ minimizing $\|\vec{p} - \vec{b}\|$.

So multiplying both sides of (**) by $(B^tB)^{-1}$ gives

$$-(B^tB)^{-1}B\vec{b} = \vec{\lambda},$$

and substituting this in (*) yields

$$\boxed{\vec{p} = \vec{b} - {}^tB(B^tB)^{-1}B\vec{b}}$$

which is our projection formula.

Exercise: Use this to project the vector $\vec{b} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ to the xy-plane, and to the plane $x - y + z = 0$.

Ans: $\vec{p} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}$ (of course) and $\frac{1}{3}\begin{pmatrix} 1 \\ 8 \\ 7 \end{pmatrix}$.

$B = (1 \quad -1 \quad 1)$

$B^tB = (1 \quad -1 \quad 1)\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = 3$

$B\vec{b} = (1 \quad -1 \quad 1)\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = 2$

$\vec{p} = \vec{b} - 2\,{}^tB = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} - \frac{2}{3}\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 8/3 \\ 7/3 \end{pmatrix}$

Now let's think about projecting to the span of a set of vectors: let $V \subset \mathbb{R}^n$ be $\mathrm{Col}(A)$, the k-dimensional column space of an $n \times k$ matrix A of rank k. To minimize $\|\vec{x} - \vec{b}\|$ with $\vec{x} \in V$ is the same as to minimize the function

$$f(\vec{y}) := \|A\vec{y} - \vec{b}\|^2 \quad \text{from } \mathbb{R}^k \to \mathbb{R}.$$

Writing $f = h \circ \vec{F}$, where $\vec{F}(\vec{y}) = A\vec{y} - \vec{b}$ and $h(\cdot) = \|\cdot\|^2$, the Chain Rule $\Rightarrow$

$$Df(\vec{y}) = Dh(A\vec{y} - \vec{b}) \cdot D\vec{F}(\vec{y})$$
$$= 2\underbrace{{}^t(A\vec{y} - \vec{b})}_{{}^t\vec{y}\,{}^tA - {}^t\vec{b}} A$$

and setting this to zero gives

$${}^t\vec{y}\,{}^tAA = {}^t\vec{b}A \Rightarrow {}^tAA\,\vec{y} = {}^tA\vec{b}$$

$$\Rightarrow \vec{y} = ({}^tAA)^{-1}\,{}^tA\vec{b} \Rightarrow \vec{x} = A\vec{y} = A({}^tAA)^{-1}\,{}^tA\vec{b}.$$

this exists for the same reason as $(B^tB)^{-1}$

Let's restate what we have derived:

Projection formulas:

$$\text{proj}_{N(B)}(\vec{b}) = \vec{b} - {}^tB(B{}^tB)^{-1}B\vec{b}$$

$$\text{proj}_{C(A)}(\vec{b}) = A({}^tAA)^{-1}{}^tA\vec{b}$$

You may notice a similarity here:

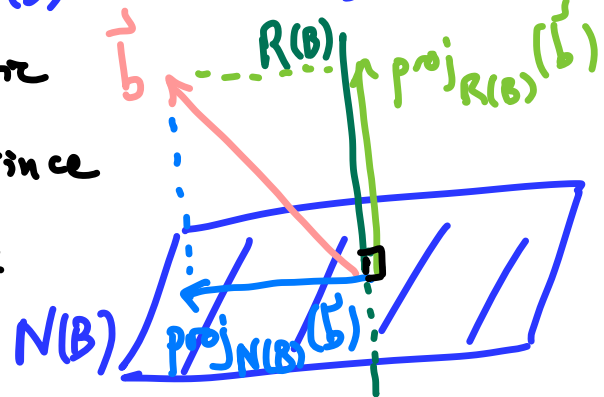
the term ${}^tB(B{}^tB)^{-1}B\vec{b}$ looks like

it should be $\text{proj}_{C({}^tB)}(\vec{b})$ $(= \text{proj}_{R(B)}(\vec{b}))$

Well, it is: $\text{proj}_{N(B)}(\vec{b}) = \vec{b} - \text{proj}_{R(B)}(\vec{b})$

or $\vec{b} = \text{proj}_{N(B)}(\vec{b}) + \text{proj}_{R(B)}(\vec{b})$

has a geometric interpretation since $R(B) = (N(B))^{\perp}$:



There is one further interpretation of the second equation, or rather its derivation:

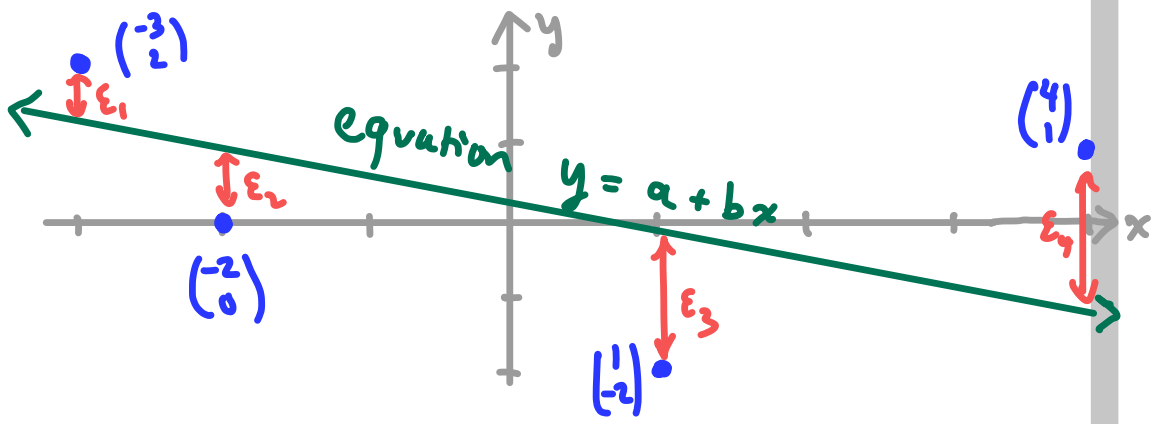
recall that we wanted to minimize $\|A\vec{x} - \vec{b}\|^2$, and the choice that did

(change name of variable)

it was $\bar{x} := ({}^tAA)^{-1}{}^tA\vec{b}$, the unique solution to the normal equation

$${}^tAA\vec{x} = {}^tA\vec{b}.$$

This is not an actual solution to $A\vec{x} = \vec{b}$, which in this situation tends not to have one, but the least squares solution.



We will take this to mean that we want to minimize the sum of squares of the vertical errors ε_i. This is the same as the length of the vector

$$\begin{pmatrix}1&-3\\1&-2\\1&1\\1&4\end{pmatrix}\begin{pmatrix}a\\b\end{pmatrix}-\begin{pmatrix}2\\0\\-2\\1\end{pmatrix}=\begin{pmatrix}a+b(-3)-2\\a+b(-2)-0\\a+b(1)-(-2)\\a+b(4)-1\end{pmatrix}=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\end{pmatrix}$$

$A \cdot \vec{x} - \vec{b}$

We compute ${}^tAA = \begin{pmatrix}1&1&1&1\\-3&-2&1&4\end{pmatrix}\begin{pmatrix}1&-3\\1&-2\\1&1\\1&4\end{pmatrix} = \begin{pmatrix}4&0\\0&14\end{pmatrix}$

and ${}^tA\vec{b} = \begin{pmatrix}1&1&1&1\\-3&-2&1&4\end{pmatrix}\begin{pmatrix}2\\0\\-2\\1\end{pmatrix} = \begin{pmatrix}1\\-1\end{pmatrix}$, then

solve the normal equation $\begin{pmatrix}4&0\\0&14\end{pmatrix}\begin{pmatrix}a\\b\end{pmatrix} = \begin{pmatrix}1\\-1\end{pmatrix}$

by $\begin{pmatrix}a\\b\end{pmatrix} = \begin{pmatrix}4&0\\0&14\end{pmatrix}^{-1}\begin{pmatrix}1\\-1\end{pmatrix} = \frac{1}{20}\begin{pmatrix}14&0\\0&4\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix} = \begin{pmatrix}1\\-1/2\end{pmatrix}$,

yielding the fitting line $y = 1 - \frac{1}{2}x$. //

There is one further interpretation of the second equation, or rather its derivation: recall that we wanted to minimize $\|A\vec{x} - \vec{b}\|^2$, and the choice that did it was $\bar{x} := ({}^tAA)^{-1}{}^tA\vec{b}$, the unique solution to the normal equation

(change name of variable)

$$\boxed{{}^tAA\,\vec{x} = {}^tA\vec{b}.}$$

This is not an actual solution to $A\vec{x} = \vec{b}$, which in this situation tends not to have one, but the least squares solution.

Example: Find a line that "best fits" the data $\begin{pmatrix}x_1\\y_1\end{pmatrix} = \begin{pmatrix}-3\\2\end{pmatrix}$, $\begin{pmatrix}x_2\\y_2\end{pmatrix} = \begin{pmatrix}-2\\0\end{pmatrix}$, $\begin{pmatrix}x_3\\y_3\end{pmatrix} = \begin{pmatrix}1\\-2\end{pmatrix}$, and $\begin{pmatrix}x_4\\y_4\end{pmatrix} = \begin{pmatrix}4\\1\end{pmatrix}$.